심리 음향 기준을 이용한 새로운 음질 개선 방법

김대경[†] · 박장식[††] · 손경식[†††]

요 약

최근에 심리 음향 기준을 이용한 스펙트럼 차감법이 제안되었다. Virag의 알고리즘에서는 기존의 방법보다 청취자가 더 편안한 음성을 들을 수 있지만 잡음에 강인한 음성활동 검출기가 필요하다. 음성활동 검출기를 필요로 하지 않는 확장 스펙트럼 차감법에서는 신호 대 잡음비가 감소함에 따라 잔여 잡음이 더욱 잘 들리게 된다. 본 논문에서는 심리 음향 기준을 이용한 스펙트럼 차감법에 Wiener 필터를 결합한 새로운 음질 개선 방법을 제안한다. 제안한 방법에서는 Wiener 필터를 사용하여 음성 구간에서도 잡음의 추정치가 계속 갱신되므로 음성 검출기가 필요 없고 마스킹 임계값에 따라 차감 파라미터를 조정하기 때문에 잔여 잡음이 거의 들리지 않게 된다. 제안된 방법에 대하여 시뮬레이션을 통하여 기존의 스펙트럼 차감법과 성능을 비교한 결과, 제안한 방법을 사용하여 개선된 음성이 기존의 방법에 비하여 청취하기에 더 편한 음질을 제공하였다.

New Speech Enhancement Method using Psychoacoustic Criteria

Kim Dae Kyung[†], Park Jang Sik[††] and Son Kyung Sik[†††]

ABSTRACT

The spectral subtraction algorithm using a criterion based on the human perception has been recently developed. The speech processed with Virag's algorithm sounds more pleasant to a human listener than those obtained by the classical methods. However, Virag's algorithm requires a robust voice activity detector (VAD). In the ESS (extended spectral subtraction) algorithm without VAD, the residual noise becomes more noticeable as the SNR decrease.

In this paper we propose a new speech enhancement method, the combination of Wiener filter and spectral subtraction based on noise masking characteristics in the human auditory system. There is no need of VAD because the noise can be successively updated even during speech activity using Wiener filter. The adjustment of the subtraction parameter based on the masking threshold makes the residual noise inaudible. The proposed method has been compared with conventional spectral subtraction algorithms. Objective and subjective evaluation of the proposed system is performed with several noise types having different time-frequency distributions. The application of objective measures, the study of the speech spectrograms, as well as subjective listening tests, confirm that the enhanced speech with proposed algorithm is more pleasant to a human listener.

1. 서 론

원하는 신호가 배경 잡음과 섞여 있는 환경에서 잡음을 제거하여 음질을 개선하는 것은 신호처리에서 중요한 분야이다[1]. 특히, 여러 가지 잡음이 섞인 음성신호는 통신시스템과 음성 인식시스템의 성능을 크게 저하시킨다. 따라서, 사용자의 편리함, 음성처리 시스템의 성능 향상 등을 위하여 음질 개선이 반드시 필요하다[2].

잡음의 단시간 진폭 스펙트럼을 추정하여 음질을 개선하는 스펙트럼 차감법(spectral subtraction)은 하나의 입력신호로부터 잡음 스펙트럼을 추정하여 입력신호 스펙트럼에서 차감함으로써 음질을 개선한다. 이 방법은 다른 방법에 비하여 상대적으로 계

[†] 동의공업대학 영상정보과 조교수
[††] 정회원, 부산대학교 전자공학과 교수
[†††] 부산대학교 전자공학과 교수

산량이 적고 구현이 용이하다는 장점이 있다[3]. 추정된 음성신호에 대한 위상은 입력신호의 위상을 그대로 사용하는데 이것은 인간의 귀가 위상 왜곡에 둔감하다는 가정에 기초한다[3].

스펙트럼 차감법의 최대 단점은 추정된 잡음 스펙트럼을 차감하여 얻어진 음성에 악기성 잔여 잡음(musical residual noise)이 발생하여 매우 귀에 거슬린다는 것이다. 악기성 잔여 잡음은 임의의 주파수에 존재하는 순음(tone)들로 이루어져 있다[4]. 잔여 잡음 제거를 위하여 많은 연구들이 수행되어 왔다. Ephraim은 MMSE (minimum mean square estimator)를 스펙트럼 차감법에 적용시키는 방법[5]을 제안하였다. McAulay는 연판정(soft-decision) 잡음 제거 필터를 사용하는 방법을 제안하였다[6]. 그러나, 이러한 알고리즘들은 낮은 신호대 잡음비(signal-to-noise ratio: SNR)를 가지는 입력신호에 대하여 성능이 좋지 못하였다[7]. 즉, 이상의 알고리즘들은 음성의 왜곡이나 잔여 잡음의 발생 없이 잡음을 제거하지는 못하였기 때문에 추정된 음성신호의 명료도가 감소하였다. 오히려 청취자들에게는 잡음이 심한 환경에서 추정된 음성이 원래의 잡음 섞인 음성보다도 더 듣기에 곤란했다[7].

최근에는 음질개선을 위하여 인간의 청각 지각에 대한 지식을 도입하여 처리하는 방식들이 제안되어 상당히 유망한 결과를 나타냈다[8]. 이러한 심리 음향 모델(psycho-acoustic model)은 이미 광대역 오디오 부호화에서 널리 사용되고 있는 인간의 청각 모델을 사용하는데 이 모델에서는 마스킹 현상에 기초를 두고 있다. 마스킹 모델은 인간의 내이(inner ear)의 주요한 분석 메커니즘인 임계 대역(critical band) 분석과 관련되어 있다. 이러한 마스킹 속성은 주로 잡음 마스킹 임계치로 모델링 된다. 청취자는 실제로 마스킹 임계치 이하로 존재하는 잡음은 있더라도 귀에는 들을 수 없게 된다[8]. 특히, Virag 은 잡음 마스킹 임계치에 따라 스펙트럼 차감 파라미터를 적절히 조정함으로써 잡음 제거, 음성 왜곡 및 잔여 잡음 발생에 대해 좋은 성능을 나타내는 방법을 제안하였다[7]. 이 방식에서는 잡음 스펙트럼 추정을 위하여 전력 스펙트럼 차감법을 사용하므로 상당히 신뢰성이 좋은 음성 활동 검출기(voice activity detector: VAD, 이하 음성검출기)가 필요하다. 그러나, 일반적으로 입력 신호의 SNR이 낮은 경우에 대하여 신뢰성 있는 음성 검출 결과를 기대하기는 어렵다고 알려져 있기 때문에 여전히 개선의 여지를 지니고 있다.

한편, Sovka는 확장 스펙트럼 차감법(extended spectral subtraction: 이하 ESS)에서 근사화된 Wiener 필터를 사용하여 음성/비음성 구간의 구분 없이 잡음 스펙트럼을 추정, 음질을 개선하는 구조를 제안하였다[9]. 음성/비음성 구간을 구별하지 않기 때문에 음성 검출기를 필요로 하지 않는 구조를 가진다는 장점은 있지만 이 알고리즘에서는 단순히 스펙트럼 차감처리만 하기 때문에 잡음 추정 값과 실제 값 사이의 차이로 인하여 여전히 잔여잡음이 발생하게 된다[10].

본 논문에서는 Virag 이 제안한 스펙트럼 차감법의 성능을 개선시키는 방법을 제안하였다. Virag 알고리즘의 잡음 스펙트럼 추정 및 스펙트럼 차감을 수행하는 부분을 근사화된 Wiener 필터로 대체하였다. 따라서, 복잡한 과정에 의해 얻어지는 음성검출기가 필요하지 않았다. 또, 음성 구간이나 비음성 구간에서 스펙트럼을 지속적으로 추정하여 잡음 스펙트럼을 입력신호 스펙트럼으로부터 차감할 수 있었기 때문에 스펙트럼 차감법의 성능을 개선시킬 수 있었다. 음질 개선 방법으로 추정된 잡음 스펙트럼을 직접적으로 차감하지 않고 추정된 잡음 스펙트럼으로부터 심리 음향 모델에 의한 잡음 마스킹 임계치를 구하여 이것을 스펙트럼 차감 파라미터 조정에 사용함으로써 잔여 잡음과 음성 왜곡이 최소가 되게 하였다.

2. 스펙트럼 차감법을 이용한 음질 개선

2.1 스펙트럼 차감법

일반적으로 스펙트럼 차감법은 그림 1과 같은 구조를 가진다. 입력 신호 x는 중첩되는 프레임으로 나누어져 처리되며 일반적으로 부가잡음 $n(k)$에 의해 오염된 음성신호 $s(k)$을 시스템의 입력으로 가지고 있다고 가정한다. 이 때 프레임 길이는 신호를 8 kHz로 샘플링을 하는 경우 128~256 샘플 사이의 값을 이용한다. 이것은 음성이 짧은 구간 동안 정적이라고 가정하였기 때문이다. 윈도우는 일반적으로 Hanning 또는 Hamming 윈도우 등을 사용한다. 시간 영역의 신호를 주파수 영역으로 변환하여 각 프레임의 입력 신호 스펙트럼을 계산하고 추정된 잡음 스펙트럼을
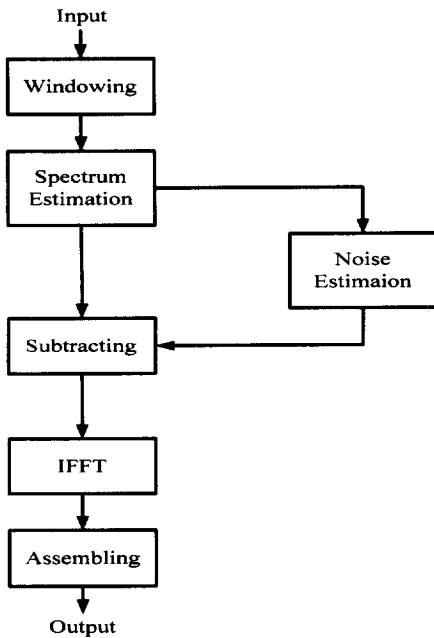
그림 1. 스펙트럼 차감법

차감하여 음성 스펙트럼이 추정된다.

추정된 음성 신호를 시간 영역으로 변환하면 원하는 시간영역에서의 음성 신호 $\hat{s}(k)$를 얻을 수 있다. 부가잡음 $n(k)$는 음성신호와 어떠한 상관도도 가지지 않으며(uncorrelated), 확률적 특성이 시간에 따라 변한다고 가정한다.

$$x(k) = s(k) + n(k) \tag{1}$$

$$X(e^{j\omega}) = S(e^{j\omega}) + N(e^{j\omega}) \tag{2}$$

$x(k)$는 시간영역에서의 입력신호, $N(e^{jw})$은 해당 프레임에서의 잡음의 스펙트럼, $S(e^{jw})$은 음성 신호의 스펙트럼이다.

주파수 영역에서 동작하는 대부분의 스펙트럼 차감법은 잡음 스펙트럼의 추정이 반드시 필요하다. 이것은 음성 검출기에 의해 판단된 비음성 구간 즉 잡음만이 있는 구간에서 잡음을 평균함으로써 가능하다. 그러나, SNR이 낮은 경우에서는 음성검출기의 신뢰성이 다소 떨어진다[12]. 이런 점이 음성 검출기를 이용하는 많은 알고리즘에서 성능을 저하시키는 심각한 원인이 된다. 또한, 음성 검출기를 이용하는 알고리즘의 경우 음성구간에서는 잡음을 갱신할 수 없거나 갱신한다고 하더라도 올바르게 갱신되지 못한다는 문제점을 가지고 있다.

음성의 전력 스펙트럼 $\hat{P}_S$를 추정하기 위해서 입력 신호의 전력 스펙트럼 $\hat{P}_X$와 잡음의 전력 스펙트럼 $\hat{P}_N$을 추정했었다. 아래 식에서 φ는 위상(phase)을 나타낸다. 잡음과 음성은 상호 독립적이라 가정하였으므로 음성 신호의 전력 추정치는 식(3), (4)와 같이 나타낼 수 있다.

$$\hat{P}_S(e^{j\omega}) = \hat{P}_X(e^{j\omega}) - \hat{P}_N(e^{j\omega}) \tag{3}$$

$$\varphi_S(e^{j\omega}) = \varphi_X(e^{j\omega}) \tag{4}$$

2.2 확장 스펙트럼 차감법(ESS)[9]

Wiener 필터를 이용한 스펙트럼 차감법은 음성의 스펙트럼의 크기를 Wiener 필터로 추정한다. 그 구조는 아래와 같고, Wiener 필터의 주파수 특성은 식(5)와 같이 표현된다.
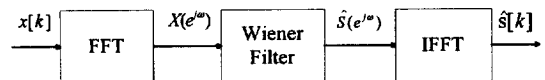
그림 2. Wiener 필터를 이용한 스펙트럼 차감법

$$H^2(e^{j\omega}) = \frac{|X(e^{j\omega})|^2 - |\bar{N}(e^{j\omega})|^2}{|X(e^{j\omega})|^2} \tag{5}$$

Sovka[9]는 ESS에서 스펙트럼 차감법과 Wiener 필터의 조합을 기초로 한 알고리즘을 제안하였다. 이 방법에서는 잡음의 통계적인 특성이 정적이거나 천천히 변화한다고 가정하고 상대적으로 통계적 특성에 빠른 변화를 가지는 것은 음성이라고 가정하였다. 이 방법에서는 음성구간에서 배경 잡음 스펙트럼 추정치를 지속적으로 갱신할 수 있다. ESS와 Wiener 필터를 이용한 알고리즘 사이에는 몇 가지 차이점이 있다. 첫째, 일반적인 Wiener필터는 비음성 구간에서 음성을 추정하는 것과는 달리 ESS에서 이용되는 근사화된 Wiener 필터는 입력 잡음의 추정치를 찾는데 이용된다. 둘째, Wiener 필터를 적용시키는 값은 입력신호를 이용하지 않고 출력신호로부터 만들어진다는 것이다. 셋째, 시스템의 구조상 피드백이 이용되므로 이 알고리즘은 필터의 계수를 지속적으로 갱신하는 적응 Wiener 필터로 볼 수 있다는 것이다. ESS의 기본구조는 그림 3과 같다. ESS에서 Wiener 필터의 주파수 응답은 식 (6)과 같다.
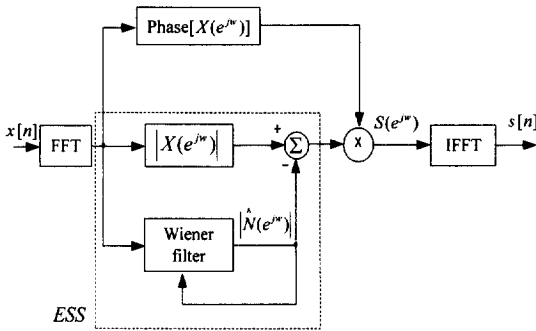
그림 3. 확장 스펙트럼 차감법(ESS)

$$H_n(e^{j\omega})=\left(\frac{|\bar{N}_{n-1}(e^{j\omega})|^2}{|\bar{N}_{n-1}(e^{j\omega})|^2-|\bar{S}_{n-1}(e^{j\omega})|^2}\right)^{1/2} \quad (6)$$

$|\bar{N}_{n-1}(e^{jw})|^2$은 잡음의 전력밀도의 추정치이고 $|\bar{S}_{n-1}(e^{jw})|^2$은 음성 전력밀도의 추정치를 나타낸다. Wiener 필터의 출력은 잡음 스펙트럼 추정치이다.

$$|\bar{N}_{n+1}(e^{i\omega})|=p|\bar{N}_n(e^{i\omega})|+(1-p)|\hat{N}_n(e^{i\omega})| \quad (7)$$

$$|\bar{S}_n(e^{i\omega})|=|X_n(e^{i\omega})|-|\bar{N}_n(e^{i\omega})| \quad (8)$$

식 (7)에서 p의 값은 전체 알고리즘의 동작에 있어서 중요한 영향을 끼친다. 만약 음성의 변화정도와 잡음의 변화정도가 잘 구분된다면 p의 값을 적절하게 정할 수 있고 결론적으로 전체 알고리즘이 좋은 성능을 나타내게 된다[9]. 통계적 특성의 완만한 변화는 $|N_{n-1}(e^{jw})|^2$에서 나타나고, 반대로 이보다 빠른 변화는 $|S_{n-1}(e^{jw})|^2$에서 나타난다.

2.3 Virag의 스펙트럼 차감법[7]

전력 스펙트럼 차감법(power spectral subtraction: PSS)은 식 (9)와 같이 표현된다.

$$|\hat{S}(e^{j\omega})|^2=\begin{cases}|X(e^{j\omega})|^2-|\hat{N}(e^{j\omega})|^2, & if\ |X(e^{j\omega})|^2>|\hat{N}(e^{j\omega})|^2\\ 0, & \text{otherwise}\end{cases} \quad (9)$$

입력 신호의 위상 성분에 대해서는 아무런 처리도 행하지 않기 때문에 만약 깨끗한 음성 스펙트럼의 크기에 입력 신호의 위상 성분을 결합하면 스펙트럼 차감법을 적용하여 얻을 수 있는 최선의 결과를 얻게 되는데 이것을 '이론적 한계(theoretical limit)'라고 부른다. 이론적 한계에 의하여 추정된 음성 파형은 식 (10)로 표시된다.

$$\hat{s}(n)=IFFT\,[|\hat{S}(e^{j\omega})|\cdot e^{j\arg X(e^{j\omega})}] \quad (10)$$

스펙트럼 차감법에서 추정된 음성은 입력 신호의 스펙트럼과 추정된 잡음의 스펙트럼에 따라 시변(time-varying) 하는 선형 시스템의 출력으로 볼 수 있는데 이것을 잡음 섞인 음성의 단시간 스펙트럼 크기와 이득 함수 $G(\omega)$의 곱으로 나타낼 수 있다. 한편, (9)식으로 표시되는 전력 스펙트럼 차감 필터는 식 (12)로 표시 가능하다.

$$\hat{S}(e^{j\omega})=G(e^{j\omega})\cdot|X(e^{j\omega})| \text{ with } \quad 0\le G(e^{j\omega})\le 1 \quad (11)$$

$$G(e^{j\omega})=\sqrt{1-\frac{|\hat{N}(e^{j\omega})|^2}{|X(e^{j\omega})|^2}} \quad (12)$$

식 (12)에서 분자가 0보다 작아지면 즉, 잡음의 스펙트럼 추정치가 입력 신호의 스펙트럼 보다 커지면 이득은 0이 된다. 이 식에서는 추정 잡음 스펙트럼보다 신호 스펙트럼이 큰 정도에 따라 입력신호의 스펙트럼 성분을 수정하고 있다. 이득함수 $G(\omega)$는 음성과 비음성 구간에서 변화한다. 즉, 음성구간에서는 입력신호를 그대로 보내고 비음성 구간에서는 완전히 입력신호를 감쇄시킨다. 이득 함수에 여러 가지 기준을 적용하여 잔여 잡음을 제거하려는 다양한 시도가 있었다[11]. 이러한 방법들은 식 (13)과 같이 일반화된 이득 함수로써 표현이 가능하다.

$$G(\omega)=\begin{cases}\left(1-\alpha\left[\frac{|\hat{N}(\omega)|}{|X(\omega)|}\right]^{\gamma_1}\right)^{\gamma_2}, & \text{if } \left[\frac{|\hat{N}(\omega)|}{|X(\omega)|}\right]^{\gamma_1}<\frac{1}{\alpha+\beta}\\ \left(\beta\left[\frac{|\hat{N}(\omega)|}{|X(\omega)|}\right]^{\gamma_1}\right)^{\gamma_2}, & \text{otherwise}\end{cases} \quad (13)$$

(13)식은 일반적인 스펙트럼 차감법 표현식으로써 기존의 방법들도 표현 가능한데 파라미터로 표시된 값을 조절하는 방법에 따라 배경 잡음과 잔여 잡음 제거, 음성 왜곡에 있어서 다른 성능을 나타낸다. 초과 차감(over-subtraction) 파라미터 α ($\alpha>1$)는 스펙트럼을 필요 이상으로 감소시키기 때문에 잔여 잡음 최고치를 감소시키지만 음성 왜곡은 증가한다[4]. 파라미터 β($0<\beta\ll 1$)는 잔여 잡음을 마스킹하기 위한 배경 잡음의 최저 한계(spectral flooring)를 결정한다. 잔여 잡음은 감소시키는 효과가 있지만 출력 음성 신호에 배경 잡음이 존재하게 된다[7]. 멱 지수

(exponent) 파라미터 γ ($\gamma=\gamma_1=1/\gamma_2$)는 입력이 그대로 전달 되게 하는 경우와 입력이 완전히 감쇠되도록 하는 경우 사이의 전환을 얼마나 급격하게 하는가를 결정한다. γ 에 따라 스펙트럼 차감법은 각각 진폭 스펙트럼 차감법 ($\gamma_1=1$, $\gamma_2=1$), 전력 스펙트럼 차감법 ($\gamma_1=2$, $\gamma_2=0.5$), Wiener 필터링 ($\gamma_1=2$, $\gamma_2=1$)과 같아지게 된다. γ 는 α 와 β보다 전체 알고리즘의 성능에 적게 영향을 미친다[4].
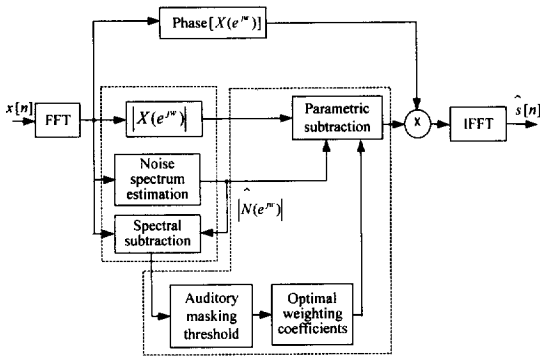
그림 4. Virag이 제안한 스펙트럼 차감법

파라미터 값을 적절히 선정하는 것은 어려운 일이지만 추정된 SNR에 따라서 파라미터 α를 적절히 변화시킨 비선형 스펙트럼 차감법에서는 개선된 결과를 나타냈다[6]. Virag은 파라미터를 시간과 주파수에 따라 변하는 잡음 마스킹 임계치에 따라서 변경하는 방법을 사용하였다. 마스킹 임계치는 주파수 영역의 마스킹 즉, 순시 마스킹 만을 고려하였다. 그림 4는 Virag에 의해서 제안된 스펙트럼 차감법을 나타낸 것으로 FFT를 이용한 입력신호의 스펙트럼 측정. 음성/비음성 검출 및 비음성 구간의 잡음 스펙트럼 추정, 잡음 마스킹 임계치 $T(\omega)$의 계산, 잡음 마스킹 임계치 $T(\omega)$에 따른 α, β의 조정, α, β 값을 식 (13)에 적용하여 개선된 음성을 얻는 과정으로 이루어져 있다. Virag은 γ 값을 실험적으로 정하고, γ는 2, 즉 γ_1은 2. γ_2는 0.5를 사용하였다.

스펙트럼 차감에 사용되는 파라미터는 시간(매 프레임 m)과 주파수(각각의 임계 대역 k)에 따라 계속적으로 바뀌는데 마스킹 임계치 $T(\omega)$에 의하여 결정된다. 파라미터 , 를 증가시키면 잔여 잡음이 감소하지만 상대적으로 더 많은 음성 왜곡이 발생하고 배경 잡음도 더 남게된다. 가장 적절한 파라미터 값은 잔여 잡음이 마스킹 임계치 이하가 되도록 하여 귀에 들리지 않게 하는 것이다. SNR이 낮은 경우에는 마스킹 임계치가 너무 낮아서 완전하게 마스킹 할 수 없으므로 합성음처럼 들리는 음성 왜곡이 발생한다. Virag은 마스킹 임계치가 높을 때는 파라미터를 최소값으로 설정하고, 마스킹 임계치가 낮은 경우에는 잔여 잡음이 귀에 들리게 되므로 차감 파라미터를 증가시키는 방법을 제안하였다.

$$\alpha_m(\omega)=F_\alpha[\alpha_{min},\alpha_{max},T(\omega)] \tag{14}$$

$$\beta_m(\omega)=F_\beta[\beta_{min},\beta_{max},T(\omega)] \tag{15}$$

매 프레임 m에 대하여, 마스킹 임계치의 최소값 ($Tm(\omega)$)은 파라미터 $\alpha_m(\omega)$, $\beta_m(\omega)$의 최대값에 해당한다. α_{min}, β_{min}과 α_{max}, β_{max}은 각각 파라미터의 최소, 최대값으로 정한 값이다. F_α 및 F_β는 마스킹 임계치의 최소값에 대해서는 잔여 잡음을 최대한 제거하기 위한 파라미터 값을 설정하고, 마스킹 임계치의 최대값에 대해서는 최소한의 잔여 잡음제거를 위한 파라미터 값 설정을 해주는 함수이다. 즉, $T(\omega)_{min}$과 $T(\omega)_{max}$가 각각 매 프레임에 갱신되는 마스킹 임계치의 최소값 및 최대값일 때 F_α 및 F_β는 식 (16), (17)과 같다.

$$F_\alpha=\alpha_{max}, \quad if\ T(\omega)=T(\omega)_{min} \tag{16}$$

$$F_\alpha=\alpha_{min}, \quad if\ T(\omega)=T(\omega)_{max} \tag{17}$$

식 (16), (17)의 두 가지 값 사이에 해당하는 함수 F_α[ ]는 $T(\omega)$의 값에 따라 보간(inter-polation)에 의하여 구해진다. 역시 F_β도 같은 방식으로 값을 정한다. 적용하는 파라미터의 값에 따라 성능이 바뀌게 되는데 α_{max}의 값을 작게하면 잔여 잡음이 증가하는 반면 음성 왜곡은 줄어들지만 β_{max}의 값을 감소시키면 잔여 잡음 뿐만 아니라 배경 잡음도 줄어들게 된다. Virag은 음성/잡음 구간 검출과 잡음 스펙트럼 추정을 위하여 에너지를 이용한 검출 알고리즘을 적용하였다[7].

3. 제안하는 스펙트럼 차감법

3.1 제안하는 스펙트럼 차감법

ESS는 음성 검출기없이 계속적으로 잡음 스펙트럼을 추정 할 수 있지만 추정된 잡음 스펙트럼을 단

순히 차감처리에만 사용함으로써 잔여잡음에 대하여 아무런 처리를 하지 않는다. Virag 이 제안한 방법은 청각 시스템의 마스킹 특성을 이용하여 잔여 잡음을 감소시키기는 효과가 있지만, 잡음 마스킹 임계치를 계산하기 위하여 잡음 스펙트럼 추정치를 필요로 한다. 따라서, 입력신호의 음성/비음성 구간을 판단할 수 있는 음성 검출기가 반드시 필요하다. 그러나 SNR이 낮은 극심한 잡음환경에서 신뢰성 있는 음성 검출기를 구현한다는 것은 매우 어려운 문제이다[12].

본 논문에서는 음질 개선 방법으로 ESS에서 사용된 근사화된 Wiener 필터를 사용하여 잡음 스펙트럼을 추정한다. 입력신호 스펙트럼에서 추정된 잡음 스펙트럼을 차감한 결과를 이용하여 잡음 마스킹 임계치를 계산하고 계산된 임계치에 따라서 매 프레임마다 임계 대역별로 파라미터를 적응시키는 Virag의 스펙트럼 차감법을 사용한다. 제안한 방법을 그림 5에 표시하였다. 첫번째 Wiener 필터는 잡음 스펙트럼을 추정하는데 사용되고 Virag의 알고리즘에 필요한 음성 검출 정보를 제공한다. '*Noise reduction*'으로 표시된 부분은 Virag의 알고리즘과 동일하다.
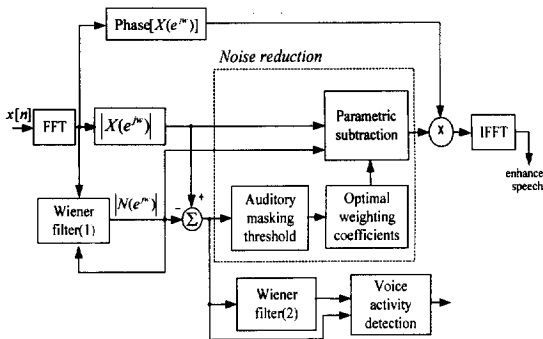
그림 5. 제안하는 스펙트럼 차감법

제안하는 알고리즘을 사용하면 계속적으로 잡음 스펙트럼을 추정하기 때문에 음성 검출기가 필요없고 잔여 잡음에 대하여 청각 특성을 고려해서 처리함으로써 더욱 개선된 음질을 기대할 수 있을 것이다. 그러나, 잡음 마스킹 임계치를 계산하고 파라미터를 조정하기 위하여 계산량이 증가한다[7].

3.2 잡음 마스킹 임계치 계산

잡음 마스킹 임계치는 귀의 주파수 선택성과 마스킹 속성을 모델링 함으로써 얻어진다. 본 논문에서 잡음 마스킹 임계치는 일반적인 음성 압축 코딩[4]에 사용되는 방법과 이를 음질개선에 적용한 Virag의 방법을 기초로 구해지는데 그림 6과 같은 과정을 거친다.
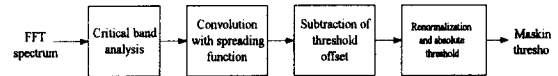
그림 6. 마스킹 임계치의 계산

임계 대역(바크)에서의 스펙트럼 분석을 통하여 해당 대역 k에서의 에너지를 구하고, 임계대역들 사이의 마스킹을 고려하기 위하여 스프레딩 함수 SF(k)[13]와 컨볼루션을 한다. 마스킹을 일으키는 신호(masker)의 순음성(tone-like)과 잡음성(noise-like) 정도에 따라 상대적인 임계치 오프셋 O(k)를 차감한다. 스프레딩 효과를 제거하기 위하여 재정규화(renormalization)을 하고 최종적으로 절대 가청 임계치와 비교하여 상대적으로 큰 값을 잡음 마스킹 임계치로 결정한다.

음질 개선 시스템에서는 단지 잡음 섞인 음성만 주어지는 환경이기 때문에 잡음 섞인 음성으로부터 잡음 마스킹 임계치를 구하는 것이 필요 하다. 잡음 섞인 음성으로부터 잡음 마스킹 임계치를 구하는 것은 매우 어렵기 때문에 전력 스펙트럼 차감법을 사용하여 추정된 음성으로부터 잡음 마스킹 임계치를 구하게 된다.

3.3 시뮬레이션 조건 및 결과 검토

각각의 알고리즘에 대하여 백색 가우시안 잡음, 자동차 잡음, 헬리콥터 잡음 환경에서의 음질 개선 성능을 시뮬레이션하였다. 시뮬레이션에서 사용된 입력신호는 깨끗한 음성신호에 인위적으로 특정한 SNR값의 배경 잡음을 더하여 만들었다. 시뮬레이션에 사용된 음성 신호는 "저 재무 관리실의 회계과 부탁합니다"라는 ETRI 전화용 데이터이며 표본화 주파수는 8 kHz이고 16비트로 양자화하였다.

자동차 잡음은 고속도로에서 80 km/h로 이동하는 소형 승용차의 운전석에서 얻어진 데이터를 이용하였다. 일반적으로 자동차 잡음은 차량 운행 환경에 따라 다르지만 입력 SNR 이 -10 dB 이상인 것으로

알려져 있다[2]. 헬리콥터 잡음은 UH-60 헬리콥터의 조종실에서 녹음한 데이터를 사용하였다. 일반적으로 헤리콥터의 비행 중 잡음은 잡음 지수 94~98dB SPL(sound pressure level)로서, 조종사와 조종사간은 물론 탑승자들 상호간에도 헤드셋(headset)이나, 조종사 헬멧에 부착되어 있는 잡음 제거 마이크를 사용하지 않으면 의사소통이 어렵다. 헬리콥터 내부에는 각seg종 무전기와 항법 장비 등 많은 전자 장비들이 있지만 소음을 제거할 수 있는 장비는 잡음제거 마이크 밖에 없으며 헬멧에 부착되어 있는 잡음 제거 마이크의 성능은 평균 감쇠 비율 24dB NNR (noise reduction rating)이고, 조종사의 잡음 제거 마이크를 통해서 나온 음성 신호는 5~10dB의 입력 신호대 잡음비(SNR)을 가진다. 시뮬레이션에 사용된 데이터는 헬멧에 부착된 잡음 제거 마이크를 사용하여 헬리콥터의 고도가 3000 feet (약 915 m), 속도가 140 knots(약 260 km/h), 온도는 섭씨15도인 조건에서 녹음하였다. 각 알고리즘에서 윈도우 길이는 256이고, 윈도우의 1/2씩 중첩되는 Hanning 윈도우를 사용하였다.

각각의 배경잡음에 대하여 ESS, Virag의 알고리즘 그리고 제안한 알고리즘의 성능을 비교하였다. 각 잡음에 대하여 식 (18)로 표시되는 입력 세그먼트 SNR(이하 SNR_{seg})을 바꿔가면서 시뮬레이션하였다.

$$SNR_{seg} = \frac{1}{M}\sum_{j=0}^{M-1} 10\log_{10}\left[\sum_{n=m_j-N+1}^{m_j} \frac{s^2(n)}{[s(n)-\hat{s}(n)]^2}\right] \quad (18)$$

각 알고리즘의 출력 음성에서 잔여 잡음 발생 정도를 평가하기 위하여 스펙트로그램(spec-trogram)을 사용하고 음성을 청취하여 그 객관적인 성능 평가 기준인 SNR_{seg} 개선 결과와 비교하였다. SNR_{seg} 개선 정도는 식 (19)와 같으며 잡음 제거 정도와 음성 왜곡 정도를 나타낸다.

$$G_{SNR} = \frac{1}{L}\sum_{m=0}^{L-1} 10\cdot\log\frac{\frac{1}{K}\sum_{k=0}^{K-1} n^2(k+Km)}{\frac{1}{K}\sum_{k=0}^{K-1}[s(k+Km)-\hat{s}(k+Km)]^2} \quad (19)$$

L은 신호의 프레임 수, N은 매 프레임 당 샘플수를 나타내며 일반적으로 15~25 msec를 하나의 프레임으로 본다.

ESS에서 p값은 0.93이고, Virag의 알고리즘의 파라미터 값은 $\alpha_{max} = 4$, $\alpha_{min} = 1$, $\beta_{max} = 0$, $\beta_{min} = 0$, $\gamma_1 = 2$, $\gamma_2 = 0.5$이고, 음성 검출을 위한 임계치로 0.7을 사용하였다. 파라미터의 값은 적용한 알고리즘이 좋은 성능을 나타내도록 여러 가지 값을 바꿔가면서 실험적으로 정하였다. 제안한 방법에서는 다른 알고리즘과 동일한 파라미터 값을 사용하였다. 부가 잡음이 백색 가우시안이고 입력 SNR_{seg}이 -9dB인 경우에 각 알고리즘의 출력 파형은 그림 7과 같다.

4000 샘플 부근과 12000~14000 샘플 사이의 묵음구간을 비교해보면 제안한 방법이 다른 방법보다 원음성에 더 유사하며 상대적으로 잔여 잡음도 줄어든 것을 알 수 있다. 5000~7000, 10000~12000 샘플사이에 있는 음성을 비교해보면 ESS 알고리즘만 적용한 경우보다 음성에서 감쇠가 적음을 알 수 있다. 한편, 같은 음성구간에 대한 Virag의 알고리즘 출력 파형은 보다 많은 잡음이 남아 있다. 출력 파형을 청취한 결과 제안한 알고리즘을 사용한 것이 잔여 잡음과 음성 왜곡의 측면에서 다른 알고리즘에 비하여 나은 결과를 나타냄을 알 수 있다. 백색 가우시안 잡음환
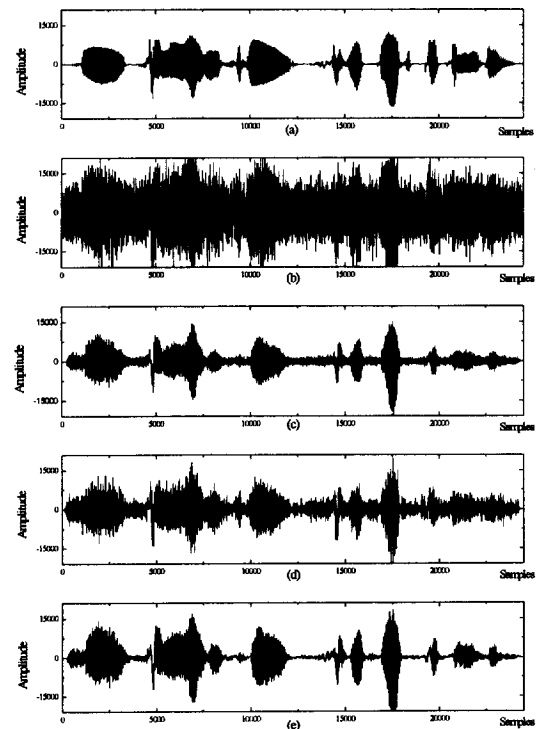
그림 7. 백색 가우시안 잡음에 대한 각 알고리즘의 처리 결과 (-9 dB)
(a) 깨끗한 음성 (b) 입력 (c) ESS (d) Virag (e) 제안한 방법

경에서 입력 SNR_{seg}이 다른 경우에도 입력 SNR_{seg}이 -9 dB 인 경우와 유사한 경향을 보였다.

부가 잡음이 자동차 잡음이고 입력 SNR_{seg}이 -9dB인 경우 제안한 방법이 다른 방법보다 묵음구간에서 잔여 잡음이 더 적었다. 음성 구간을 비교해보면 ESS를 적용한 경우보다 음성왜곡이 더 적었다. 그러나, Virag의 알고리즘에서는 여전히 잡음이 많이 남아 있었다. 청취 결과에서도 제안한 알고리즘이 다른 알고리즘에 비하여 더 나은 결과를 나타내었다. 자동차 잡음에서 입력 SNR_{seg}이 다른 경우에도 제안한 방법이 다른 방법에 비하여 우수한 경향을 나타내었다. 부가 잡음이 헬리콥터 조종석 내부의 잡음인 경우에도 각 알고리즘이 자동차 잡음에서의 성능과 유사한 경향을 나타냈다.

각각의 배경 잡음과 잡음 레벨에 대하여 기존의 알고리즘들과 제안한 알고리즘을 적용했을 때 SNR_{seg}이 개선되는 정도를 그림 8에 나타내었다. 제안한 알고리즘을 백색 가우시안 잡음에 적용한 경우에 입력 SNR_{seg}의 값에 따라 출력 SNR_{seg}이 기존의 방법보다 45dB 높게 나타났으며 입력 SNR_{seg}이 커질수록 그 차이가 커졌다. 자동차 잡음의 경우에 제안한 방법은 기존의 방법보다 2~3dB정도 높은 SNR_{seg}을 나타냈다.

헬리콥터 잡음에 대하여, 제안한 방법은 기존의 방법보다 SNR_{seg}이 4~6dB 정도 높게 나타났다. 따라서 제안한 방법이 기존의 방법에 비하여 전반적으로 더 나은 결과를 나타냄을 알 수 있다. 한편, 입력 SNR_{seg}이 높은 경우에는 아무런 처리를 하지 않은 입력 음성이 음질 개선 알고리즘을 적용한 것보다 SNR_{seg}이 더 높게 나타나는데 이것은 청취 결과와는 상반되는 것으로, SNR_{seg}이 직접적으로 인간의 청취와 관련되어 있지는 않다는 것을 의미한다.

입력 SNR_{seg}이 -9dB인 백색 가우시안 잡음인 경우에 대하여 각 알고리즘의 출력 신호 스펙트로그램을 그림 9에 표시하였다. 스펙트로그램에 사용된 윈도우는 256 샘플의 Hanning 윈도우를 128 샘플씩 겹쳐지게 하여 사용하였다. (b)에 ESS를 적용한 음성 스펙트로그램을 나타내었는데 100~125 프레임 정도에서 나타나는 스펙트럼이 평탄한 부분이 거의 복원 되지 못하고 전 주파수 영역에서 고르게 잡음이 분포하는 것을 알 수 있다. (c)에 나타난 Virag의 알고리즘을 적용한 출력 음성에 대해서도 유사한 경향
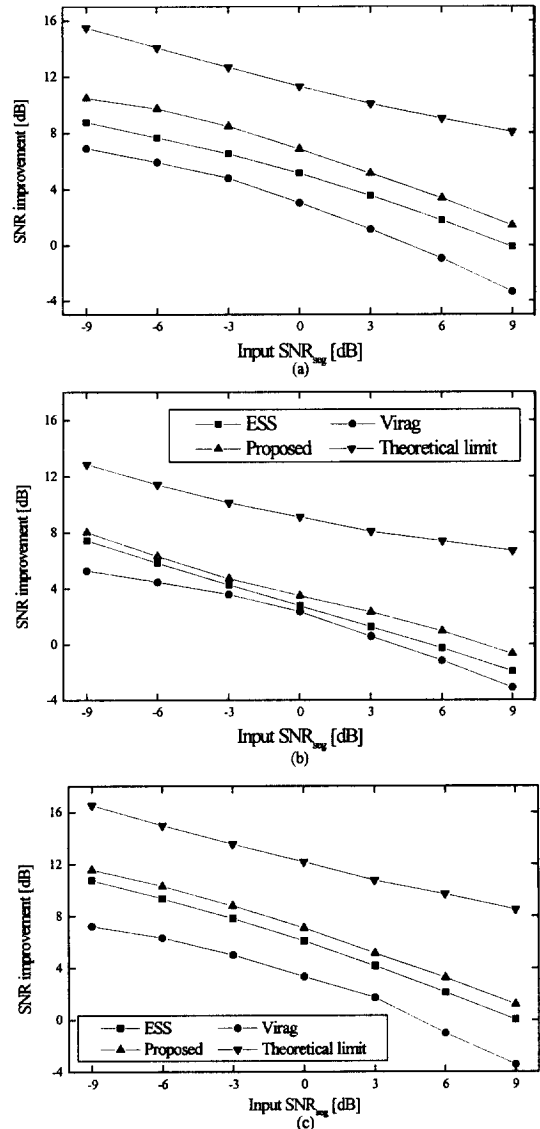
그림 8. 배경잡음과 잡음 레벨에 따른 각 알고리즘의 SNR 개선
(a) 백색 가우시안 잡음 (b) 자동차 잡음 (c) 헬리콥터 잡음

을 볼 수 있다.

그림 (b),(c)에서 공통적으로 잡음 감소 효과는 볼 수 있지만 중간주파수 이상의 영역에서 계속적으로 잔여 잡음이 발생하는 것을 볼 수 있다. (f)에서 제안한 방법을 적용한 경우 원음성 (a)와 유사하게 잡음이 상당히 많이 감소되어 100~125프레임 정도에서 나타나는 스펙트럼이 평탄한 부분이 잘 드러나 보임을 알 수 있다. 75 프레임 부근에서 출력음성의 스펙

그림 9. 음성 스펙트로그램 (백색 가우시안 잡음)
(a) 깨끗한 음성 (b) ESS
(c) Virag (d) 제안한 방법

트럼이 깨끗한 음성의 스펙트럼과 거의 유사하게 나타난다. 잔여 잡음은 기존의 방법에 비하여 훨씬 적게 발생하는 것을 알 수 있다.

자동차 잡음에 대하여 각 알고리즘 출력의 스펙트로그램을 비교해보면 ESS를 적용한 출력에서는 중간 주파수 대역의 스펙트럼이 평탄한 부분이 거의 복원 되지 못하였다. Virag의 알고리즘을 적용한 경우에는 출력에서는 오히려 잔여 잡음이 ESS의 경우보다 더 많이 발생하였다. 제안한 방법을 적용한 경우 묵음부분이 잘 드러나 보이고 잔여 잡음이 기존의 알고리즘에 비하여 더 적게 발생하였다. 그러나, 잔여 잡음을 억제하는 과정에서 음성 스펙트럼에 약간의 왜곡이 발생하였다. 헬리콥터 잡음의 경우에는 ESS를 적용한 출력에서 저주파대역의 잡음이 많이 감소하지만 3kHz에서 아직도 잡음이 많이 남아 있었다. Virag 알고리즘을 적용한 경우, 잔여잡음이 ESS와 유사하게 발생하였고 3kHz에서 잡음이 더 많이 남아있었다. 제안한 방법을 사용한 경우 잡음 성분이 더 제거되었고 고주파 영역까지 잡음이 현저하게 감소하였다. 기존의 알고리즘에 비하여 잔여 잡음도 훨씬 적게 발생하는 것을 확인할 수 있었다.

4. 결 론

최근에 인간의 청각 지각에 대한 지식을 스펙트럼 차감법에 적용하는 방식들이 제안되었다. Virag은 잡음 마스킹 임계치를 이용하여 파라미터를 조정하는 스펙트럼 차감법을 제안하였다. 이 방법은 다른 방법에 비하여 상대적으로 계산량이 적고 구현이 용이하지만 잡음 스펙트럼 추정을 위하여 음성/비음성 구간을 구분하므로 신뢰성 있는 음성 검출기가 필요하게 된다. 그러나, 입력 신호가 SNR이 낮을 경우 신뢰성 있는 음성 검출을 기대하기 어렵기 때문에 개선된 음성에 잔여 잡음이 발생하거나 음성 왜곡이 생겨서 추정된 음성 신호의 명료도가 감소하였다.

한편, Sovka는 근사화된 Wiener 필터를 사용하여 음성/비음성 구간의 구분없이 잡음 스펙트럼을 추정하여 스펙트럼 차감에 사용하는 ESS을 제안하였다. 음성/비음성 구간을 구별하지 않기 때문에 음성 검출기가 필요 없다는 장점은 있지만 이 알고리즘에서는 단순히 스펙트럼 차감처리만 하기 때문에 잡음 스펙트럼 추정 오차로 인하여 잔여잡음이 계속적으

로 발생하였다.

본 논문에서는 음질 개선을 위하여 ESS과 Virag의 방법을 결합한 구조를 제안하였다. 제안한 방법에서 근사화된 Wiener 필터를 사용하여 음성/비음성 구간에 상관없이 잡음 스펙트럼을 계속적으로 추정하였다. 입력신호 스펙트럼에서 추정된 잡음 스펙트럼을 차감하여 잡음 마스킹 임계치를 계산하고 그 값에 따라서 매 프레임마다 파라미터를 적응시키는 Virag의 스펙트럼 차감법을 사용하였다.

깨끗한 음성에 각각 백색 가우시안 잡음, 자동차 잡음, 헬리콥터 잡음을 입력 SNR_{seg}이 -9 dB~9 dB가 되도록 하여 ESS, Virag의 방법 및 제안한 알고리즘을 적용하여 음질 개선 성능을 비교하였다. 제안한 방법을 적용한 경우, 잡음의 종류와 레벨에 무관하게 출력 파형이 기존의 방법을 적용한 것보다 음성 파형이 적게 감쇠되었고 잔여 잡음이 더 적어졌다. SNR_{seg}은 기존의 방법에 비하여 상대적으로 백색 가우시안 잡음에서 4~5 dB, 자동차 잡음에서 2~3 dB, 헬리콥터 잡음에서 4~6 dB 정도 좋게 나타났다. 배경 잡음이 자동차 잡음일 때 SNR 개선과 청취결과가 다른 잡음에 비하여 떨어지는 것은 자동차 잡음의 스펙트럼이 음성과 유사한 스펙트럼 특성을 가지기 때문이다. 잔여 잡음의 발생정도를 나타내는 스펙트로그램에서도 제안한 방법에서 잔여 잡음이 더 적게 발생하였다.

참 고 문 헌

[1] William A. Harrison, J. S. Lim, and Elliot Singer, "A new application of adaptive noise cancellation," *IEEE. Trans. ASSP.* vol. ASSP-34, no. 1, pp. 21-27, Feb. 1986.

[2] 김대경, "심리 음향 기준을 이용한 음질 개선과 음성 활동 검출," 부산대학교 박사학위 논문, Feb. 2000.

[3] R. Martin, "Spectral subtraction based on minimum statistics," in *Proceedings of EUSIPCO-94, 7th European Signal Processing Conference*, pp. 1182-1185, 1994

[4] Saeed V. Vaseghi, "*Advanced Signal Processing and Digital Noise Reduction*", Wiley & Teubner, 1996, pp. 242-259.

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. on ASSP*, vol. ASSP-33, no. 2, Apr. 1985.

[6] R. J. McAulay and M. L. Malpass, "Speech enhancement using soft decision noise suppression filter," *IEEE Trans*, vol. ASSP-28, pp. 137-145, Apr. 1980.

[7] N. Virag, "Speech enhancement based on masking properties of the auditory system," in *Proc. IEEE ICASSP*, Detroit, MI, pp. 796-799, May. 1995

[8] Stefan Gustafsson, Peter Jax and Peter Vary, "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," in *Proc. IEEE ICASSP*, pp. 314-317, 1998.

[9] Pavel Sovka, Peter Pollak and Jan Kybic, Extended spectral subtraction, "*European Conference on Signal Processing and Communication*," Trieste, pp. 454-459, Sept. 1996.

[10] O. Cappe, "Elimination of the musical noise phenomenon with Ephraim and Malah noise suppressor," *IEEE Trans. Speech and Audio processing*, vol. 2, No. 2, pp. 345-349, Apr. 1994.

[11] S. F. Boll and D. C. Pulsipher, "Suppression of acoustic noise in speech using two microphone ANC," *IEEE Trans. on ASSP*, vol. ASSP-28, no. 6, pp. 155-157, Dec. 1980.

[12] M. Shozakai, S. Nakamura, K. Shikano, "Robust speech recognition in car environment," in *Proc. IEEE ICASSP*. vol. 1. pp. 269-272, 1998.

[13] P. Noll, "Wideband speech and audio coding," *IEEE Communication Magazine*, vol. 26, pp. 34-44, Nov. 1993.

김 대 경

1992년 2월 부산대학교 전자공학과(공학사)
1995년 2월 부산대학교 전자공학과(공학석사)
2000년 2월 부산대학교 전자공학과(공학박사)
1998년 3월~현재 동의공업대학 영상정보과 조교수
관심분야 : 디지털 신호처리, 음성 신호처리, 멀티미디어 통신

손 경 식

1973년 2월 부산대학교 전자공학과 졸업(공학사)
1977년 8월 부산대학교 전자공학과(공학석사)
1991년 8월 경북대학교 전자공학과 졸업(공학박사)
1979년~현재 부산대학교 전자공학과 교수
관심분야 : 디지털 신호처리, 적응 신호처리, 음성 및 음향 신호처리

박 장 식

1992년 2월 부산대학교 전자공학과(공학사)
1994년 2월 부산대학교 전자공학과(공학석사)
1999년 2월 부산대학교 전자공학과(공학박사)
1997년 3월~현재 동의공업대학 영상정보과 조교수
관심분야 : 음성 및 음향 신호 처리, 멀티미디어 통신, 입체음향